IOP Conf. Series: Materials Science and Engineering **966** (2020) 012084 doi:10.1088/1757-899X/966/1/012084

Informative modeling of subjective reality for intellectual anthropomorphic robots

A G Dolganov[1] and K Y Letnev[1, 2]

[1] Ural Federal University named after the First President of Russia B N Yeltsin, 19, Mira str., Yekaterinburg, 620000, Russia

E-mail: [2] ptmir@inbox.ru

Abstract. This research is aimed at solving the problem of safeguarding robot operation by modeling the subjective reality of an intellectual anthropomorphic robot. Analysis has been carried out on trends in designing advanced robot control systems and methods of improving the safety of robot usage. The following conclusions are drawn: the risk of causing damage to human life, health or property increases when people interact with robots; it is necessary to combine neural network control systems with expert control systems in order to enhance the intellect of a robot and raise the level of trust held by humans towards robots; developing norms and regulations, designing collaborative robots, drawing visualization diagrams for safety zones cannot really provide the depth of robot socialization sufficient for the human society; an informative model of the subjective reality could be integrated into an intellectual anthropomorphic robot to provide a safer 'human-robot' interaction. Such an approach should result in achieving the greatest trust of humans and their safety due to anthropomorphic conversion of both the external design of a robot and its internal control structure. By its nature, the research is interdisciplinary – it is run in the fields of artificial intelligence and philosophy of subjective reality.

1. Introduction

The national program 'Digital Economy of the Russian Federation' defines robotics as a comprehensive technology of the modern manufacturing [1]. The World Economic Forum (WEF) has assessed the world-wide proportion of robotized (even to a degree) manufacturing to be about 29 %. An analytical review of the world robotics market published by Sberbank of Russia shows that robots would continue to spread, and the total number of industrial robots used in production should have grown two times by 2021 in comparison with 2019 [2]. Analysts of the International Federation of Robotics (IFR) predict that the market capacity of industrial robotics would have exceeded $210 billion by 2022 (if software part is included into the calculation) [3].

Nowadays, one of the trends in development of robotics is to increase the proportion of intellectual anthropomorphic robots. The risk of harming people would rise with the improvement of intellectual capacities and anthropomorphic features of the robot.

Today, most active integration is observed in neural network control systems (NNCS). Many researchers note limitations of such NNCS-approaches. In particular, Leif Jentoff, general director of RightHand Robotics, says that 'neural network is not a solution ready for all problems. The technology just seems to be powerful, but it is not really true' [2]. In the near future, one can expect a transition towards cooperative usage of NNCSs and expert control systems (ECS) in robots. It could be suggested that NNCSs would be utilized to solve relatively simple intellectual problems in the

framework of weak artificial intelligence (AI) concept, while ECS – to solve complex intellectual problems in the framework of strong AI concept.

The principle of NNCS operation is based on modeling the inner workings of the nervous system within the human mind. An ECS simulates the informative process of making decisions, human thinking. It follows that, in practical terms, these types of AI systems differ in their ability and quality of reproducing the process of decision making and explaining the logic behind those decisions to humans. 'Important trend related to the AI progress is an all-growing demand for the explanation of how an AI takes one decision or another. With the growth of AI capabilities, it is trusted with more decisions and actions – and they become more critical and complex. If we fail to trust the AI as a whole, development of newer applications for autonomous intellectual robots could slow down' [2].

One cannot argue that an NNCS, as an information system of a robot, is non-transparent to the human, thus remaining a closed information system (black box) in those cases when one needs to decode its workings using a human-transparent verbal language. And it leads to natural mistrust which people feel towards robots and their protection while interacting with them.

2. Existing trends in solving problems of robot operation safety

Solving problems of robot operation safety is realized in several directions. In Russia, there are a few national standards on robotics [4–7]. These standards provide regulations for safe designs of robots. They are important but not enough since danger could come not only from a mechatronic robot system but also from an intellectual robot control system (RCS).

Recently, significant progress has been made in developing so-called cobots. Developers of cobots stress that they are safe for the human [8]. But cobots today are characterized by a weak intellect, simple functions, limited transport mobility.

Two world-wide leaders in the market of robotics, a Japanese FANUC Corporation and German KUKA Robotics Corporation, claim higher safety of their robots and, in particular, control systems for operators. For example, FANUC offers a software product called DCS – an intellectual integrated software to safeguard operators, robots and tools: 'Using FANUC iPendant Touch operators are able to visualize defined safety zones and confirm this from a 3D perspective in front of the robot cell' [9].

Analysis of world trends in robotics shows that capabilities of robots grow at high rate, problems they solve become more complex, interaction with humans intensifies. For example, soon demand will rise not just for robotic mechanisms performing few simple and repeated tasks but for robots with a specific profession: robot-vendor, robot-mechanic and so on. These expectations of robot-market consumers actualize the problem of social intellect in robots. Not only should a robot be able to communicate with people, but it also has to be intellectually secure, that is to make socially responsible decisions which would ensure protection of human life, health, property based on accepted social norms [2].

The study made within the framework of the national program 'Digital Economy of the Russian Federation' on legislation of robotics points out at possible principles of regulation in robot design and engineering: 1) prohibition of damage on the part of a robot; 2) robot as a human assistant and not his substitute; 3) humans have a bit of control over a robot at all times; 4) robot has to continuously record and store in its 'black box' the information about conditions of its operation and actions; 5) robot which physically interacts with humans and is not under their direct control has to have a 'red button' function of instantaneous or emergency switching off on demand; 6) design of a robot has to be secure, which includes protection of users and third-party individuals in an emergency situation and necessary testing and certification of the robot; 7) dealing with a robot should always imply respecting human dignity [1, 2].

A 'Moral Machine' developed by Massachusetts Institute of Technology demonstrates how difficult it is to simulate the human ethics within a robot [10]. Finally, the growth of intellectual robot capabilities could develop into the problem of an unpredictable strong AI whose capacities would exceed those of a natural intellect [2].

3. Solving the problem of robot operation safety by modelling subjective reality

Analysis of existing solutions on the problem of robot operation safety allows to formulate main requirements for a control system of intellectual anthropomorphic robots with the goal of securing the 'human-robot' interaction. An RCS should satisfy the following basic requirements: 1) to simulate the human 'control system' in order for a person to be able to understand the mechanism of robot control as thoroughly as he understands his own 'control mechanism'; 2) to be centralized in order to eliminate contradictions and control conflicts; to achieve this, the center of robot control should simulate the 'control center' of a human; 3) to be human-transparent in terms of information, thus eliminating risks connected to random processes influencing the decision making during the whole period of robot operation; 4) to self-identify not only while communicating with humans but constantly and continuously during the whole period of its functioning; 5) to be a self-organized system, that is to minimize failures of a robot occurring due to external and internal (for example, related to internal contradictions within the RCS between a robot's goal and methods of achieving it) factors; 6) to identify an individual with whom a robot interacts as being identical (equal) to itself in order to enable protection of life, health and property of the individual with regards to his biological and social characteristics, simulate relations of partnership and exclude relations of competition and confrontation; 7) to restore memories about past processes of information conversion, operation conditions in order to enable a possibility of reproducing and analyzing them on demand of a human-operator; 8) to explain decisions and actions of a robot on demand of a human, thus simulating human reflection over his own thinking and behavior; 9) to possess a mechanism of stopping and emergency switch-off for a robot without causing damage to a human and his property; 10) to simulate both praxiological and axiological goals of a human, in particular, ideals of human dignity, rights, freedoms and variety of cultures.

To realize the requirements cited above and solve the problem of securing the 'human-robot' interaction, the authors of this research offer an approach based on modeling the subjective reality (SR) of the human within the ECS of an intellectual anthropomorphic robot. The necessity of such an approach could be justified by the following statements:

1) Robot ECS allows to implement elements of a strong AI, that is such an AI which is compatible with a natural intellect.

2) Robot ECS can be based on production rules. Those productions are characterized by features useful for an RCS: a) modularity, which enables combining smaller fragments of knowledge into various algorithms of decision making, thus increasing RCS flexibility; b) addition of new rules independent of existing ones, which extends intellectual capacities of the RCS; c) substitution of older rules with newer ones independent of other rules, which increases longevity of the RCS; d) transparency, which allows to get from the RCS an explanation of decisions made (obtaining answers to such questions like 'how?' and 'why?') and increase the trust of humans towards robots.

3) Robot ECS can be efficiently combined with a NNCS, with the NNCS in charge of a weak AI (at the level of empiric perception), and the ECS responsible for a strong AI (at the level of reasoning and deduction).

4) High rates of robotization in manufacturing and service allow to predict appearance of a strong-AI robot in the nearest future.

5) Anthropomorphic conversion of the robot – which, in the 20-th century, is initiated by demand for visualization of a robot in the form of organoprojection – nowadays turns into the necessity of adapting the robot to the workspace of a human, integrating it into the manufacturing and service environment designed for humans. For example, experts of the Schunk company opinionate that robots should adapt to the human and 'fit in with its interior' [11]. Therefore, further anthropomorphic conversion of robots has be related to simulation of the human control system within the system of reasoning and deduction in the robot ECS.

6) Human SR is an efficient control system which is characterized by its economy, responsiveness, centralization, integrity and autonomy. According to a famous philosopher and AI expert D I Dubrovsky, the human subjective reality consists of 'conscious psychic states of an individual

which verify the fact of his own existence to him' [12]. The human SR as a type of the information process comes from the evolution and self-organization of a biological system. The information in that SR refers directly to the objective reality of a human and himself as a subject. 'The whole variety of SR phenomena occurring both consequently and simultaneously are managed and controlled by our own I…' [12]. Via his SR, a person interacts with the objective reality. The SR sets values of human relations, defines the human activity. Specifics of the SR lie in the fact that the information is expressed in a human-transparent form (figurative, verbal), and the human is in direct control of it. In the opinion of D I Dubrovsky, 'right now we only know two SR types – human and animal ones, but in theory other types are conceivable… It is true even for just theoretically conceivable products of information technologies and robotics in accordance with the principle of information invariance of a storage medium to physical properties and the statement (proved by A Turing) about isofunctionality of systems resulting from that principle'[12]. The SR is capable of connecting information systems of different types: for example, 'human-robot', 'robot-robot', 'robot-human-robot'. Another feature of the SR is its ability to run its own monitoring (self-representation of information about itself) [13].

7) It is known that the center (core) of the human SR is his 'I'. According to the definition given by a well-known specialist in the field of 'subject-object' relations V A Lectorsky, 'I is something which controls my body, the authority enabling free will of making decisions and in charge of their realization and consequences' [2]. A famous thesis of R Descartes 'cogito ergo sum' is interpreted as impossibility of a doubt in the existence of human 'I' [14]. Uniqueness of 'I' and its property of duality is stressed by representatives of the classical philosophic rationality – I Kant [15], JG Fichte [16], E Husserl [17].

The significance of the 'I-robot' model for solving the problem of safeguarding the 'human-robot' interaction is in providing: 1) wholeness, integrity of a robot: in the RCS, the 'I-robot' model can function as a center of robot control; 2) community, consolidation of robot-agent network; 3) self-identification of a robot and its identification of both other robot-agents and people; 4) self-preservation of a robot in time and space; 5) communication of a robot with other robot-agents and people; 6) purposefulness of robot behavior.

A classical notion about subjective reality factors in the property of human 'I' duality. From the viewpoint of the classical rationality, the SR can be divided into two main regions – 'internal I' and 'external I'. It follows that the 'I-robot' model should also allow for this SR duality. In its turn, that duality manifests itself in dual SR properties: constancy, subjectivity, heuristicity, generality of 'internal I' and variability, objectivity, algorithmizability and singularity of 'external I' [18]. But it is also necessary to note differences of these SR regions. While 'internal I' is the core of the human (robot) SR, 'external I' – its periphery. Confession should be made on one important limitation of the 'I-robot' model: in accordance with concepts of classical rationality, it is the fact that 'internal I' cannot be objectified since it is a subject. At the same time, 'internal I' preserves its heuristic cognoscibility. This limitation of the 'I-robot' model could be considered as a proof of the fact that an isomorphic model of 'internal I' is impossible in principle: the robot could never substitute the human to the whole extent. This situation corresponds to one of the requirements for safety in 'human-robot' relations. Yet, the problem of securing the 'human-robot' interaction perseveres.

Informative modeling of the SR for an intellectual anthropomorphic robot could be achieved by explicating productions rules in the ECS knowledge representation system. It is known that the human SR is a self-organizing system in itself, which is sorted out in the action of reflection over thinking. Therefore, the SR model of an intellectual anthropomorphic robot could be developed by representing, in the form of ECS production rules, bits of reasoning over the question: who is an 'I-robot'?

Consider a fragment from top levels in the tree of reasoning within the ECS knowledge representation system as a variant of the SR informative model for an intellectual anthropomorphic robot:

1. If 'I' am a robot, then who is an 'I-robot'? →

1.1. If I am a robot, then I am a robot body →

1.1.1. If I am a robot body, then I am a set *F* of physical properties of the robot body →

1.1.2. If I am a set *F* of physical properties of the robot body, then I am a variable, objective, algorithmizable, singular robot body →

1.1.2.1. If I am a variable robot body, then I am a value of the function for an argument *a* from a set *A* of states and positions for the robot body →

1.1.2.2. If I am an objective robot body, then I am a value of the function for an argument *b* from a set *B* of parameters and indicators for designs of a mechatronic robot-body →

1.1.2.3. If I am an algorithmizable robot body, then I am a value of the function for an argument *c* from a set *C* of procedures and operations to change states and positions of the robot body →

1.1.2.4. If I am a singular robot body, then I am a value of the function for an argument *d* from a set *D* for individual codes of robot-agent bodies →…

1.2. If I am a robot, then I am a subject of control for robot-agent bodies →

1.2.1. If I am a subject of control for robot-agent bodies, then I am a set *P* of ideal properties for the subject of control for robot-agent bodies →

1.2.2. If I am a set *P* of ideal properties for a subject of control for robot-agent bodies, then I am a constant, subjective, heuristic, general subject of control for robot-agent bodies →

1.2.2.1. If I am a constant subject of control for robot-agent bodies, then I am a subject of control for robot-agent bodies identical to myself in space and time →

1.2.2.2. If I am a subjective subject of control for robot-agent bodies, then I am an ideal (virtual, cloud-based) subject of control for robot-agent bodies →

1.2.2.3. If I am a heuristic subject of control for robot-agent bodies, then I am a value of the function for an argument *e* from a set *E* for production rules in a robot-body ECS →

1.2.2.4. If I am a general subject of control for robot-agent bodies, then I am a collective subject of control (universe *U*) for robot-agent bodies →…

An ideal body of the robot SR can be implemented in the information environment of cloud computing services which are used for machine learning of robots. For example, experts estimate that ‘one can expect mass influx of cloud-based services which would unite, aggregate and grant access to collective knowledge for robots… Even now, developers experiment with clouds: for example, Amazon with their AWS IoT Greengrass AWS IoT Greengrass or Google with Cloud Robotics (a cloud computing platform for robot control). And soon we shall see mass production of such solutions…’ [2].

It is necessary to clarify that the ‘subjective subject’ term used above is logically justified since a ‘subject’ in itself has no unambiguous definition in literature, while even being often confused with an ‘object’ (subject is identified with object). Therefore, the ‘subjective subject’ stresses non-objectivity of a subject in its physical form, which is a fundamental constraint in modeling an SR.

Thus, modeling the subjective reality of an intellectual anthropomorphic robot can secure the ‘human-robot’ interaction due to the following factors: 1) simulation of a human SR in the robot ECS as a result of prolonged biological evolution and improvement; 2) centralization of the RCS both for a single robot and a group of robot-agents; 3) properties of the robot-SR core: constancy (ensuring stability in safeguarding robot operation), subjectivity (enabling direct and fast access to a mechatronic robot-control system by utilizing knowledge stored in the ECS), heuristicity (providing transfer of heuristic knowledge from a human to a robot, thus increasing efficiency of a robot in information processing), generality (ensuring unification of the RCS within the network of robot-agents).

4. Conclusion

1. Analysis of world robotics market shows a constant growth of production volumes for industrial and service robots. In relation to this, one should expect higher risks of getting life, health and property of humans damaged from interaction with robots.

2. Advance in development of control systems for robots should lead to combining neural networks and expert systems since it is going to strengthen the robot intellect and equip robots with a capability of explaining their actions and decisions on human demand. As a result, such combination of these systems should increase the trust of humans towards robots.

3. Nowadays, the problem of safeguarding operation of robots is solved by writing regulations, developing designs of collaborative robots, producing visualization diagrams for safety zones. But all those solutions cannot reach the necessary level of socialization for robots in the human society, especially in the sphere of population service.

4. This research suggests a solution to the problem of securing operation of robots by modeling subjective reality of an intellectual anthropomorphic robot. With that goal in mind, main requirements were defined for the robot control system; justification was made of an approach which is based on modeling the subjective reality of the human within an expect control system; functions of an 'I-robot' model and its fundamental constraint were determined; a fragment from top levels in the tree of reasoning within the system of knowledge representation for an expert control system was presented as a variant of the informative model of subjective reality for an intellectual anthropomorphic robot; factors were formulated, which are useful for adhering to safety of 'human-robot' interaction while modeling the subjective reality of an intellectual anthropomorphic robot.

References

[1] The national program 'Digital Economy of the Russian Federation'. Available online: http://government.ru/rugovclassifier/614/events/ (accessed on 01.06.2020)
[2] Sberbank. Analytic review of the world robotics market, 2019. Available online: http://www.sberbank.ru/common/img/uploaded/pdf/sberbank_robotics_review_2019_17.07.2019_m.pdf (accessed on 01.06.2020)
[3] International Federation of Robotics - Representing the global robotics industry, 2018. Available online: https://ifr.org/downloads/press2018/WR_Presentation_Industry_and_Service_Robots_rev_5_12_18.pdf (accessed on 01.06.2020)
[4] Russian state standard GOST R ISO 8373-2014 Robots and robot devices. Terminology and definitions
[5] Russian state standard GOST R 60.0.0.2-2016 Robots and robot devices. Classification
[6] Russian state standard GOST R 60.2.2.1-2016 Robots and robot devices. Requirements for robot safety and maintenance
[7] Russian state standard GOST R 60.0.2.1-2016 Robots and robot devices. General requirements for safety
[8] *Collaborative revolution: what to expect and whether to fear.* Available online: https://robo-hunter.com/news/kollaborativnaya-revolyciya-chego-jdat-i-stoit-li-opasatsya9045 (accessed on 01.06.2020)
[9] *FANUC Corporation*. Available online: https://www.fanuc.eu/ru/ (accessed on 01.06.2020)
[10] *Moral Machine*. Available online: http://moralmachine.mit.edu/ (accessed on 01.06.2020)
[11] *Schunk Company*. Available online: https://schunk.com/ru_ru/kompania/o-kompanii/ (accessed on 01.06.2020)
[12] *Encyclopedia of Epistemology and Philosophy of Science 2009* (Moscow: 'Canon-plus' ROOI 'Rehabilitation') p 1248
[13] Nichols S and Stich S 2003 *How to read your own mind: A cognitive theory of self-consciousness Consciousness.* New philosophical perspectives ed Q Smith and A Jokic (Oxford: Oxford University Press)
[14] Descartes R 1950 *Discourse on Method and Metaphysical Meditations Selected works* (Moscow: GIPL) p 250
[15] Kant I 1965 *Prolegomena to Any Future Metaphysics Collected works* (in 6 vols) vol 4 part 1 (Moscow: GIPL) p 320
[16] Fichte J G 1993 *Collected works (in 2 vols)* (Saint-Petersburg: Mifril) p 1485
[17] Husserl E 2001 *Cartesian Meditations* (Saint-Petersburg: Science) p 264
[18] *On self-examination as a method of human self-identification.* Available online: https://sites.google.com/site/problemyznania/home (accessed on 01.06.2020)